DOCUMENT RESUME

ED 450 778 IR 058 034

AUTHOR Mannerheim, Johan
TITLE The WWW and Our Digital Heritage--The New Preservation Tasks of the Library Community.
PUB DATE 2000-08-00
NOTE 9p.; In: IFLA Council and General Conference: Conference Proceedings (66th, Jerusalem, Israel, August 13-18, 2000); see IR 057 981.
AVAILABLE FROM For full text: http://www.ifla.org/IV/ifla66/papers/158-157e.htm.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Access to Information; *Electronic Publishing; Foreign Countries; International Programs; *Library Role; Library Services; *Preservation; *World Wide Web
IDENTIFIERS *Digital Data; Metadata; Web Sites

ABSTRACT

This paper discusses the role of libraries in the preservation of World Wide Web publications. Topics addressed include: (1) the scope of Web preservation, including examples of projects that illustrate comprehensive and selective approaches; (2) the responsibility of Web preservation, including placing the responsibility on publishers and other institutions, the national approach, and the international approach; (3) challenges of Web preservation, including information retrieval, the short life span of Web publications, and lack of a legal framework for Web preservation and access; (4) preservation of digital information, including different digital formats and conversion of document files to readable formats; (5) the present situation of Web archiving in the world, including examples of several projects. (MES)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 450 778

IFLANET Search Contacts
International Federation of Library Associations and Institutions
Annual Conference

66th IFLA Council and General Conference

Jerusalem, Israel, 13-18 August

Conference Proceedings

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

A.L. Van Wesemael

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Code Number: 158-157-E
Division Number: VI
Professional Group: Preservation and Conservation
Joint Meeting with:
Meeting Number: 157
Simultaneous Interpretation: *No*

The WWW and our digital heritage - the new preservation tasks of the library community

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Johan Mannerheim
Division of IT, The Royal Library of Sweden, National Library of SwedenStockholm, Sweden

Paper

Traditionally the development of libraries is linked with the development of text distribution. The Alexandria library and other early libraries in Africa, Europe and the Middle East as well as early Chinese libraries like the imperial ones, and European medieval monastery and university libraries were all engaged in the acquisition and copying of manuscripts as well as giving access to them. Later the focus of libraries turned towards printed books, and still later periodicals, newspapers and reports became important forms of information for some of them.

Through deposit legislation the national libraries got the special assignment to collect, preserve and give access to everything or almost everything published in each country, that is to those printed texts which have been spread to the public.

To be able to fulfil their different tasks preservation of collections has been essential to most libraries from their very beginnings. Preservation precautions and conservation methods have been developed to take care of manuscripts on silk, papyrus, parchment, palm leafs and paper, printed books mostly on paper of different qualities and to preserve their bindings and covers or produce new ones for better protection.

The development of computers has had a growing impact on both the publishing industry and libraries since the nineteen seventies. The publishers were provided with better and better tools to produce traditional books and periodicals on paper. But from the late eighties you will find a small number (compared to printed

IR058034

BEST COPY AVAILABLE

publications) of publications using floppy disks, CD-ROM and DVD as carriers of digital information. The libraries were also early users of computers. They made it possible for them to start a revolution in cataloguing methods and routines. On preservation and conservation, however, information technology so far has only had marginal effects.

Since 1994 information technology has developed very fast in communication. Especially the Internet and WorldWideWeb have risen from an insecure alternative to dominance. The possibility to publish an illustrated article almost instantly and make it available all over the world to anybody equipped with a modem and a web browser, and to do that at a very low cost compared with the cost of producing traditional printed publications, has led to an explosive growth of web publishing.

According to a study by Steve Lawrence and Lee Giles the number of public web pages could be estimated to 800 million in February 1999. So it is safe to say that by now there are well over one milliard web pages published on the Internet. Although the number of printed publications per year does not seem to be seriously affected yet, there certainly has been a tremendous shift of balance between digital publications and paper publications within a few years time.

Among these web publications there is lots of trash, but also material of great value, both to us today and to coming generations of people, who might have a historic interest in what we are doing. It certainly is an important task of the library community to collect part or all of this and preserve it to safeguard access to it for hundreds and thousands of years! Web publications

To be able to preserve web publications it is necessary to know the construction of the WorldWideWeb and some definitions used to describe it.

The WorldWideWeb is a way of viewing pieces of information located in all different places on the Internet as if they were one large indexed document by using hypertext and multimedia techniques. This means that in a way it is impossible to preserve single publications completely because they have links pointing to other documents, which in their turn link to others.

A recent study, Graph structure in the web by Andrei Broder, Ravi Kumar and others, gives a more elaborate picture in the form of a "bow tie", separating the "strongly connected components" from looser ones. In this group of strongly connected web pages it should be possible to get from any page to any other in a small number of clicks. In their investigation of more than 200 million pages 28 percent belonged to this group, the core of the web. They also defined three other groups which happened to be of about the same size, 21-22 percent. From the IN-group you can reach any core page in a small number of clicks, but the links are not going back to them. Probably this group mainly consists of relatively new pages. The OUT-group pages can easily be reached from any of the core pages, but they are not linking back to the core. The third group, called the tendrils, is not connected to the core but to some of the other pages. Completely unconnected in this study were only 8 percent of the pages.

Hypertext mark-up language, HTML, is the language of web pages that makes this possible. An HTML file contains text, layout information, links to images and multimedia and external links (or pointers) to other pages or sites. A web browser is the software you use to download and view web pages. It interprets the HTML and image files and brings them together on your screen.

A web site is a collection of web pages run by an information provider on a server connected to the Internet. It is identified by an address called URL (uniform

resource location) up to the first "/". Here is an example:
www.kb.se/ All page addresses on that site begin with these characters (or to be precise with http://www.kb.se/).

A web page or, as some prefer to call it, a web document is what you download and view on the computer screen after asking for an address (URL):
www.kb.se/oppet.htm
www.kb.se/ENG/kbstart.htm In its simplest form it consists of an HTML file only, but usually the page is illustrated or has some logotype on it. These images are also files with their own addresses. They are automatically fetched by the HTML code. If the page has frames each frame is a separate file.

The home page of a site is the page you get if you address the site without specifying a certain page. Usually it mainly consists of links leading into the web page structure of the site. The larger sites are often constructed with a hierarchy of different catalogues and subcatalogues containing the actual web pages:
www.kb.se/Bus/DC/metadata.htm

The domain concept is needed when you discuss the over all structure of the URL system and web addresses. Highest in rank are the top-level domains. Some are national like .se (Sweden), .uk (United Kingdom), .il (Israel) and .nu (the small island Niue). Others are international like .com (business), .org (organisations).

The top-level domains are each handled by an organisation. You must apply for a domain name for your institution or company and pay to get it registered. I might look like kb.se/ telia.com/ telia.se/ The domain name owner might create subdomains for organisational or other reasons like libris.kb.se/ sbi.kb.se/. Eventually the servers, the machines containing the web sites are named, so the complete site address might look like
www.kb.se/ unow.kb.se/ www.telia.com/ www.libris.kb.se/.

The scope of web preservation

In today's projects you will find two main approaches to the scope of web preservation..

The comprehensive one is represented by the Swedish Kulturarw3 Project, by Brewster Kahle's Internet Archive and, more recently, by the Finnish EVA Project. The scope is to collect everything published on the Internet. These projects are collecting millions of documents. The selective approach is represented by the PANDORA Project of the National Library of Australia and EPPP (Electronic Publications Pilot Project) of the National Library of Canada. The scope is to collect important publications that can be made accessible at once. They are "only" collecting thousands of documents. There is also the Danish way. They changed their deposit law, so that from 1998 web publications matching certain narrow criteria should be given notice of to their national library. The result so far has not exceeded one thousand publications.

An argument for being selective is that you should not spend your limited resources on preserving lots of trash. However, doing an intelligent selection is difficult and researchers in the future will criticise our choices. Even if we try our very best, important digital information will get lost.

Computer storage is getting cheaper and cheaper, while the cost of personnel is not. It might seem a paradox, but it is a fact that the selective projects use more staff than the comprehensive ones.

If selection is made in the indexing process, and not in the collecting process, we have at least saved the publications and the inevitable mistakes we will do, when we select publications for cataloguing and indexing, can be corrected in the future.

The responsibility of web preservation

Who should preserve the digital publications? There are at least three approaches to this problem. One is to put the immediate responsibility on publishers and other institutions as was advocated in the USA by the Task Force on Archiving of Digital Information in 1996. The second is the national approach exemplified by Denmark and by the Australian, Canadian, Finnish and Swedish projects. The third is the international one represented by the Internet Archive.

Long-term preservation should be undertaken by long-term institutions with stable financing that lasts for hundreds of years. To give the task to the national library in each country, widening its responsibility for printed publications to include digital publications, based on rewriting the deposit law, seems to be a good solution for many countries. Collection and preservation is best done at one institution with good recourses, while indexing and selection might be done in co-operation with other institutions.

The institutional approach is not so stable. It also combines badly with automatic, comprehensive collecting of web publications, as each publisher and institution will find their own solution for preservation of their own publications. Links pointing to resources on other sites will not function.

The interactive character of the web pages with links to other pages, regardless of national boundaries, speaks for the international approach. But there seems to be a long road to go before it would be possible to create an international institution for web archiving with long-term stable financing. It seems more realistic to start co-operation between national web archives not only to exchange experience and provide each other with support, but to create a forum for raising questions of standards, exchange formats, communication between the archives, etc.

Waiting for a permanent solution, which seems close in Sweden and Finland, but so far fairly distant in most other countries, institutions, companies and individuals have to rely on themselves if they want to preserve their old web pages. In countries with the selective approach they have to consider if they are content with the selection.

Some challenges of web preservation

One interesting feature of web publishing is that it is so cheap and easy to accomplish. You are not bound to the traditional publishing industry and its routines, obstacles and time-consuming methods. Therefore many people who only were consumers of printed publications now are becoming involved in the production of publications on the Internet. Many become creators of text and images. A new profession of graphical designers of web pages has emerged. The web publishers are so many that there is no statistics over them yet. The number of web sites could however be used as a rough estimate. The figure is to high because many web publishers have several sites. But it is also too low as web hotels may have many publishers on the same site. In Sweden there are today more than 60 000 web sites, about twenty times as many as the traditional publishers. So it is evident that if you want to preserve their publications, you need automatic means of collecting.

If you search for a tree on the Internet today, you will get the whole forest as an

answer. In the long list presented, you are lucky if you find a relevant hit on page seven. This problem will not be less and the list will not be shorter in a historic web archive. Cataloguing, even if it is done at a minimum level, can hardly be accomplished for more than some per mil of the total number of web pages. Therefore, it is important to promote the use of metadata, in order to help and encourage the producers to make their own cataloguing and put that onto the page.

After years of discussion, it seems that the Internet community rallies around the metadata format Dublin Core. The Royal Library promotes metadata by meetings and by information on the web, by having a template for Dublin Core creation, and by encouraging other actors also to provide Dublin Core templates.

Automatic indexing and cataloguing might also be used more for digitised material in the future. There are some rather promising projects going on.

Another feature of web publications is their short life. The average life length has long been estimated to be three or four months. A small internal study made at the Royal Library shows that after a year only one fifth of the Swedish web pages were left completely unchanged, that is they still had the same check sum. About half of the pages had vanished. Their addresses (URLs) did not exist any more. The remaining pages had been changed, maybe just corrected in some detail, maybe filled with completely new contents. The check sum method does not make a difference. A manual study of a small sample suggests that most changes might be fairly marginal and would not affect indexing or cataloguing at all. There is a need for further investigation and analysis into this matter. Anyway it is clear that if you are not quick enough in collecting, many web pages will be lost forever.

Another problem is the lack of a legal framework for web preservation and access in most countries. There is not only the need of revising the deposit law, also the copyright and privacy legislation might be in conflict with web preservation and reasonable studies in the web archive.

Preservation of digital information

I will now discuss long-term preservation and access of digital information in general, of which web publications constitute one subset. The amount of digital information created is increasing drastically. The time when word processors and economy systems were tools to create written or printed documents is gone. Now more and more information is primarily digital. It might be in a text format like MS Word, HTML or XML, in an image format like TIFF or JPEG, in some kind of data base or in a more specialised system. Today not only printouts but also printed reports should often be regarded as secondary forms, which are used to spread the information or a selection of it on paper. Different digital formats like HTML, PDF and reports in Excel could as well be secondary forms to spread the information on intranets or the Internet. But for long-term preservation, most institutions and companies still stick to paper and in some case microfilm, when they are not closing their eyes to the problem.

For long-term preservation of the web it is evident that this is not possible to do. A web page could not be preserved on paper or microfilm because the hypertext and multimedia techniques embedded will get lost and can never be retrieved again. The links will point into the air. Only a shadow of the web pages will be preserved, if their functionality vanishes.

Web preservation is such a clear case. There are so many web publications out there. They are part of our cultural heritage. Their life length is short. They will be lost, if we do not do something. We have to build infrastructures to preserve them

in digital form to preserve contents as well as appearance and functionality. We have to build human and technical infrastructures for long-term preservation of digital information. And this is not so complicated.

It is the digital file that should be preserved, not the media carrying it. In this respect it is easier than paper conservation. If you make a copy of a digital file, you get the original once again if it is properly done.

The problem is the software, to interpret the digits you have saved, when the programmes that created the digits are getting outdated and the systems are shifted. To keep the digital information alive and accessible in the future you have to take care of the document files by converting them to readable formats or by applying emulation software on the original files, a software which functions in the current IT environment. Since the collections of web publications will grow large the methods applied must be as automatic as possible.

So just as you need professionals to take care of manuscripts and printed books by preservation and conservation, there will emerge experts specialised in taking care of digital documents and publications. Today tape robot archives supplemented by large hard disk arrays seem to be the cheapest solution, tomorrow other media will emerge. All file formats must be treated individually and decisions must be made about how and when the documents should be converted or emulation programmes implemented. By conversion the original file must be kept to make emulation from the authentic file possible in the future.

Long-term preservation of web publications is in principle not different from long-term preservation of any other digital information. Maybe the situation is a little easier because over 95 percent of the files, HTML and image files, are in standardised formats. So the prospect of having software reading them in the future is better than in other areas using proprietary software.

When you are building an infrastructure for long-term preservation of web publications you can as well use it for any other digital information. Most libraries have a few floppy disks and some more CD-ROMs in their collection as parts of combined publications or as independent publications, many of them not readable any more. Some libraries and especially the national libraries should investigate these resources and decide if it is worth the effort to convert some or all of them and put them online using the long-term preservation infrastructure. There are also other objects on the Internet than web pages like discussion lists and FTP collections that might be considered as publications. The Norwegian National Library and The Internet Archive have both collected Usenet documents.

I will take another example from the perspective of a research library: what happens to the manuscripts of today? For centuries, The Royal Library has collected personal archives of authors and other persons related to the publication and production of books. These are frequently used sources for studies in literature, art, history and other academic disciplines. Today, the corresponding material is in the author's PC till she or he buys a new computer, when most of it gets lost. Therefore, on the initiative of the author and professor Sven Lindqvist, a member of the library board, the IT division and Manuscripts have just started a joint project to find ways of preserving digital personal archives. Such an archive might include different versions of texts reflecting the creative process, as well as e-mail correspondence and research material collected by the author.

A last example, the results of different digitisation projects constitute another kind of digital collections worthy of long-term preservation if the quality is good enough. Some of them might have images in dying formats that should be

converted. Others might use outdated database software for stand-alone machines and need a data base conversion to become available online. Especially digital images made to reduce the use of fragile originals are worth professional handling. In many libraries the results of digitisation will constitute a large share of their digital collections.

The present situation of web archiving in the world

The first web archiving projects were Electronic Publication Pilot Project (EPPP) of The National Library of Canada which started in June 1994 and the Australian Pandora Project from June 1996. Both national libraries now collect web publications on permanent bases. They have chosen a selective approach and are cataloguing their electronic collections.

In October 1996 Internet Archive started collecting web pages from all over the world in large scale. The archive is a non-profit organisation situated in San Fransisco and founded by Brewster Kahle. Till now they have collected about 1000 million web pages.

In September 1996 the Swedish national library started the Kulturarw3 project aiming at complete collection of Swedish web pages. The regular collecting started in April 1997. Sevens "snapshots" has been taken so far comprising about 35 million web pages. A fairly informal working group Nordic Web Archive started in 1997.

In June 1997 Finland started the EVA project with complete web preservation as one of its goals. They joined the European project NEDLIB, Networked European Deposit Library, which among many other things has the development of web archiving software on their programme. It started in January 1998.

In Denmark they started by revising the deposit law and started a very selective collection accordingly in January 1998.

In France the government have initiated studies preparing their preservation of web publications and several institutions in other countries have also shown their interest in different ways.

I hope that this session will inspire libraries all over the world to raise the question of preserving the web in their countries. There is certainly need for much more co-operation on this issue. My vision is a net of national libraries all archiving their countries' web publications, so you can follow a link on a page in one archive to a page in another just as in the living WorldWideWeb.

Act on Copyright Deposit of Published Works. Translation of Act No. 423 of 10 June 1997. File No. 1997.2301-1 of the Danish Ministry of Culture. (http://www.kb.dk/kb/dept/nbo/da/pligtafl/pligt-en.htm)

Andrei Broder, Ravi Kumar, Farzin Maghoul and others, Graph structure in the web, Proceedings 9th WWW, 2000. (http://www.almaden.ibm.com/cs/k53/www9.final/)

Internet Archive. (http://www.archive.org/)

Kulturarw3. (http://kulturarw3.kb.se/html/kulturarw3.eng.html)

Steve Lawrence and Lee Giles, Accessibility of information on the web, Nature 400(1999):107-9.

National Library of Canada Electronic Collection. (http://collection.nlc-bnc.ca/e-coll-e/index-e.htm)

Networked European Deposit Library (Nedlib) (http://www.kb.nl/coop/nedlib/)

Project EVA. (http://linnea.helsinki.fi/eva/english.html)

PANDORA Project. (http://pandora.nla.gov.au/)

Preserving Digital Information, Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group, Inc. May 1, 1996. (http://www.rlg.org/ArchTF/)

Latest Revision: *August 22, 2000*

Copyright © 1995-2000
International Federation of Library Associations and Institutions
www.ifla.org